漢文の返り点推定における返り点有無の情報の有効性について

情報科学科　牛場 智子　　指導教員：山村 毅

1　はじめに

漢文とは，古典中国語で書かれた文章のことであり，日本の歴史や文学を知る上で，とても重要な存在である．我々日本人にとって，中国語は外国語の一種であり，漢文をそのまま理解することはできない．そこで，返り点と呼ばれる「漢文を日本語の語順に直すための記号」が生み出され，漢文に関する文法的知識がなくても，漢文を読み理解することができるようになった．よって，返り点は，漢文を理解する上で重要な情報であるといえる．

先行研究 [1] では，CRF(条件付き確率場) を用いて，白文[*1]に適切な返り点を推定する手法を提案した．そこでは，漢字 5 つからなる漢詩データ (以下，韻文テキストと呼ぶ) については，52.9%．物語の漢文データ (以下，散文テキストと呼ぶ) については，34.2%．さらにこの 2 つを混ぜ合わせた漢文データ (以下、混合テキストと呼ぶ) については，42.3% の精度であった．

先行研究では，品詞情報を素性として，返り点推定を行っていたが，本研究では，品詞情報を用いることなく，白文から得られる情報のみを利用し，白文の返り点推定を行うことを目的としている．

2　新たな素性の提案

返り点を推定するために，「返り点有無の情報」を素性として用いることを考える．

白文から「返り点有無の情報」は得られないため，あらかじめ「返り点有無の情報」が与えられたと仮定し，その情報を用いて返り点推定を行う．用いた分類器は，先行研究と同様に CRF++[2]，テキストデータは先行研究の韻文テキストと散文テキストである．

用いた素性は，「単語そのもの」「文頭からの順番」「文頭か否か」「文末か否か、または読点の手前か」「返読文字か否か」「再読文字か否か」「置き字か否か」「推定したい単語の前後の単語」の 8 つである．表 1 に，「返り点有無の情報」を用いた場合 (素性有り) と用いなかった場合 (素性無し) の結果を示す．

表 1　返り点有無の情報を用いた返り点推定の結果

	文章正解数 / 文章数	精度 (%)
韻文 (素性無し)	126 / 410	30.7
韻文 (素性有り)	361 / 410	88.0
散文 (素性無し)	128 / 424	30.2
散文 (素性有り)	224 / 424	52.8

素性無しの場合において，有意水準 5% の適合度検定により有意差を求めた．韻文では 145 文章以上，散文では 148 文章以上正解していれば素性無しより有意であるといえる．

表 1 を見ると，素性有りの場合は，韻文では 361 文章，散文では 224 文章と正しく返り点を推定することができた．いずれも，素性無しの結果を大幅に上回り，「返り点有無の情報」を用いた返り点推定は有効であるといえる．しかし，ここで問題となってくるのは，「返り点有無の情報」をどのように得るのかということであり，これが新たに本研究の課題となる．

[*1] 白文：返り点のついていない漢字のみの文のこと

3　実験と考察

3.1　実験と結果

ここからは，白文から返り点を推定するのではなく，白文から返り点の有無を推定することを考える．

分類器には，CRF++ とナイーブベイズ分類器の 2 つを使用した．用いた素性は，CRF++ は先ほど述べた 8 つの素性をもとに，素性の前後関係の情報も加えた．ナイーブベイズ分類器は「単語そのもの (以下，ユニグラムと呼ぶ)」「単語そのものと次の単語 (以下，バイグラムと呼ぶ)」を別々に用いて実験を行った．

推定結果を表 2 に示す．

表 2　返り点有無の推定の結果

	×正解数	F 値	○正解数	F 値
韻文 (CRF)	1270/1477	87.1	396/565	67.8
韻文 (ユニグラム)	1375/1477	85.9	218/565	49.3
韻文 (バイグラム)	1298/1477	84.4	265/565	52.5
散文 (CRF)	2165/2338	89.4	624/966	70.8
散文 (ユニグラム)	2141/2338	86.7	504/966	60.5
散文 (バイグラム)	2065/2338	86.7	608/966	65.8

∗「×」は返り点が付かない，「○」は返り点が付くことを表す．

3.2　考察

表 2 から分かるように，ナイーブベイズ分類器より CRF++ を用いた方が良い結果を得られた．ナイーブベイズでは，一つの素性のみを用いて推定を行ったため，推定するための情報少なく，良い結果が得られなかったと考えられる．

韻文と散文，どちらにおいても，「○」の F 値は 60 前後と「×」より 20 近く低いことがわかる．これは，「×」に比べ「○」の推定結果が悪いことを示している．

この結果を用いて返り点の推定を行うためには，「○」についても F 値 90 ほどの結果を得る必要があるだろう．

4　まとめ

本研究では，「返り点有無の情報」を用いて返り点推定を行うことで，精度を上げることができることを示した．また，CRF++ とナイーブベイズ分類器を用いて「返り点有無」の推定を行った．しかし，まだ良い結果が得られず，返り点推定の素性として用いるには不十分であると考えられる．今後の課題として，素性が返り点有無の推定に与える影響を分析し，有効な素性を見つける必要がある．また，これは二値分類問題であるため，サポートベクターマシン (SVM) を用いて推定を行うと良いのではないかと考えられる．

参考文献

[1] 牛場智子，佐藤綾花，山村毅: ”CRF を用いた漢文の返り点推定”, 電気・電子・情報関係学会東海支部連合大会講演論文集，L4-4, 中京大学, 2014

[2] Taku Kudo: "CRF++: Yet Another CRF toolkit", http://crfpp.googlecode.com/svn/trunk/doc/index.html